The importance of having data-sets.

Ronald Dekker
Delft University of Technology

The importance of having data-sets

Ronald Dekker

Delft University of Technology, The Netherlands
r.dekker@library.tudelft.nl

Abstract

Much scientific research is based on the gathering and analysis of measurement data. Scientific data-sets are, at least, intermediate results in many scientific research projects. For some time data-sets weren't even published and even if they were published it was mostly as a (not re-usable) by-product of the publication. But an interesting phenomenon might be observed here: data-sets (often in combination with models and parameters) are becoming more important themselves and can sometimes be seen as the primary intellectual output of the research. Publishing and preserving data-sets should therefore seriously be considered. This will especially be the case if the data cannot be reproduced (as they result from unique events) and will be necessary in the future for longitudinal research or to test or check future insights.

The rationale behind the importance of data-sets can be summarized as follows:

- Verification of publications
  (results= analysis + data)
- Longitudinal research
  (long periods, meta-research)
- Interdisciplinary use of data
  (reuse/innovation)
- Valorisation
  (get new projects based on data-set ownership)

The increasing importance of data-sets may lead to their emancipation into an essential component of an institutions scientific infrastructure. Networks of institutional repositories might form a basic building block for such an infrastructure.

In this paper three essential processes with respect to the importance for datasets will be explained:

- capturing data: collaborate with researchers during the data gathering and analysis phase of their research,
- publish data: facilitate the publication of research data on it's own or related to a publication,
- preserving data: take care of the long term preservation of digital datasets.

In this paper the case of the DARELUX (Data Archiving River Environment LUXembourg) project will be used to provide a relation between theory and practice. In the DARELUX project the preservation of hydrology data-sets using XML containers is being investigated and implemented. The OAI-PMH protocol for meta-data harvesting is being used to provide resource discovery.

An archive like the DARELUX repository can only exist with its users on top of mind. We adopted the CCSDS term `designated community` to explore how we can meet our users' needs. From this notion stems our conviction that an archive like the DARELUX repository should be an integral part of the research community.

Keywords: Data Archive, Digital Preservation.

1 The scientific workflow

The process leading to a scientific publication can be described by the following basic workflow:

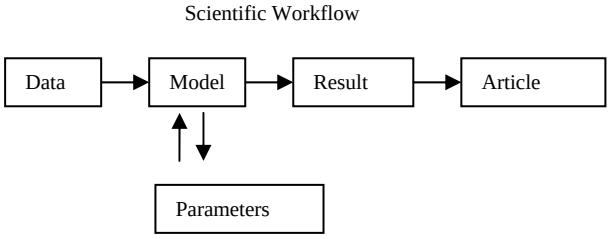

Table 1: Process leading to a scientific publication

The analysis of data in this workflow is an important step in the scientific research. In general, this analysis is conducted using different types of models and parameter sets. After publishing the final research results in an article, data-sets, models and parameters are, intentional or not, more often than not, being discarded.

There are however arguments why these data might be worthwhile to publish and preserve:

Verification. A scientific publication has to be self-descriptive so it can be repeated by others providing the same results.

In reality it is hard to meet this requirement of self-descriptiveness, due to constrictions in the length of publications and the use of complex models. Often the analysis leading to the publication is described within the paper, but the data and the models itself are not archived.

Longitudinal research. Sometimes research takes place over a long period of time. Taking care of the preservation aspects of the data that are related to this type of research is of utmost importance since the long term availability is an essential part of the research,

Interdisciplinary use of data. (re) Using data for a different purpose than that it was intentionally gathered for might lead to new insights, and enhance innovation.

Valorisation. The ownership of valuable datasets may create income or enable the acquisition of new research projects.

2 About the research

The research that privides tha datasets for the DARELUX project is carried out in the valley of the Maisbich and the Huelerbach in Luxembourg. It is a longitudinal hydrological research to enhance discharge models that describe the relation between rainfall and discharge witch of course holds an importance in relation to water management, environmental management and flood control.

Three essential processes with respect to the processing of the datasets will be explained:

- capturing data: collaborate with researchers during the data gathering and analysis phase of their research,
- publish data: facilitate the publication of research data on it's own or related to a publication,
- preserving data: take care of the long term preservation of digital datasets.

2.1 Capturing data

Since the gathering and analysis of data is the heart of the research that is being carried out it is essential that the archive is very near to the researcher. On the other hand, the data has to be put in the archive as soon as possible to take care of its availability over time. While storing data in the archive relevant meta-data should be added. In accordance with the OAIS reference model, the archive stores reference, context, provenance and fixity meta-data.

During the research project, the archive will facilitate the researchers and, in fact, be embedded in the actual research. Such an environment may grow into what is sometimes called a collaboratory. Researchers and data archivist are together involved in storing the data, adding meta-data and preparing the datasets for long term preservation.

In the DARELUX [1] project data will be collected from field sensors, the raw data is enriched by quality control procedures (e.g. checks on the completeness of the time series) and the addition of metadata. It is also possible to add relevant data from other sources to the archive.

The researchers (data producers) are actively involved in quality control and meta-data addition since they possess the knowledge about the data. It is therefore essential that storing the data in the archive is done at the moment when all relevant knowledge is still available and not when the project has ended and the knowledge and attention of the researchers has shifted elsewhere.

2.2 Categorizing users (the designated community) of the repository

The Consultative Committee for Space Data Systems (CCSDS) [2] defines the so called designated community as: "an identified group of potential consumers who should be able to understand a particular set of information. The designated community may be composed of multiple user communities". Since this designated community is the target group of users the archive aims its services at, its definition should be made clear.

For the ease of reasoning we will divide our designated community into three groups: primary, secondary and tertiary users.

We will define primary users as users who act as users and producers of the data at the same time. Primary users are actually building-up the archive as a result of their research. This data is used and for analysis. Without primary users no archive will start to exist. It is this group who has the major interest in setting up the archive to support their research. The expertise of these researchers is of eminent importance to provide quality assurance during the ingest of data into the archive. Primary users will have an interest in the operational archive for (at least) the duration of their research project.

Secondary users are not actually participating in the research of the primary users, but they possess data that is of interest to the primary users. At the same time, mutate mutandis; the data stored in the archive by the primary users might be of interest to the secondary users. Storing data of secondary users in the archive should be undertaken on the presumption that this adds value to the archive as perceived by the primary users. Storing data of secondary users in the archive should only be undertaken under the prerequisite that conditions for use and re-use of this data are clear and agreed on.

Tertiary users are users who might take an interest in the data stored in the archive in a near or further future. This is, at least at the moment when the archive being built, a largely unknown group.

2.3 Publishing data-sets

Much scientific research is based on the gathering and analysis of measurement data. Scientific data-sets are, at least, intermediate results in many scientific research projects. For some time data-sets weren't even published and even if they were published it was mostly as a (not re-usable) by-product of the publication. But an interesting phenomenon might be observed here: data-sets (often in combination with models and parameters) are becoming more important themselves and can sometimes be seen as the primary intellectual output of the research. Publishing and preserving data-sets should therefore seriously be considered. This will especially be the case if the data cannot be reproduced (as they result from unique events) and will be necessary in the future for longitudinal research or to test or check future insights.

The DARELUX repository is being used as a primary publishing platform for the data-sets. The OIA-PMH protocol is being used for resource discovery.

Data publication today suffers from several flaws in the publication process [3]. Due to economic and practical constraints data is not published in journals. Since data is not citable there is little merit for the author in publishing the data. One of the reasons that inhibit the citability of data is the often transient character of web locations (URLs). So to make data citable there is a need for the use of persistent identifiers.

The German National Library of Science and Technology (TIB) has completed the pilot phase of a project to install DOIs (digital object identifiers) that continuously retrieve scientific datasets. [4]

Based on these results the DARELUX project will start publishing datasets in combination with the article.

To become citable datasets the assignment of persistent identifiers is essential. But that will not be all. Apart from persistent identifiers, they also need to be stored in trusted long-term archives. This calls for specific action taken with respect to the long term preservation of data-sets.

2.4 Preservation of digital data-sets

While taking care of the preservation of data-sets, three major issues have to be addressed:

- the physical decay of storage media
- the loss of (descriptive) meta-data: inability to retrieve information
- the loss of "rendering" functionality caused by the inability to run old software on new computers and operating systems

In the DARELUX project the concept and technology that was designed and developed in the e-Archive [5,6] project that dealt with digital preservation is being used.

Preventing physical degradation is managed by replication the data on a new medium when need be (an adequate average might be every 10 years). We store information objects as simple bit streams. To prevent irretrievability we store meta-data and information objects inseparable together in a so called XML container. The original bit stream of the information object is always stored in the archive. Emulation or conversion strategies are being used to provide perpetual rendering.

In order to avoid the data becoming obsolete, a digital archive must choose a strategy for the format of the bit streams. The three options are:

1. **A priori conversion to a generic format**. 'Generic' here means that every file has the same format after conversion and the resulting format is independent from any application, operation system or hardware platform. Every data-set is converted at the ingest process.
2. **A posteriori conversion to a generic format**. When the original format will soon become obsolete, the data-sets are converted to the generic format. This way, conversion only occurs if it is absolutely necessary and at that moment it can be reconsidered whether the data-sets will still be needed in the future.
3. **A posteriori incremental conversion**. This boils down to the idea that the data is converted to a new version of the old format. An advantage is that viewer-development will be easier, as filters to convert to the new version may already exist. A disadvantage is that the data will have to be converted again and again, possibly until eternity.

The outcome of this choice has no consequences for the original bit streams. As was found in the e-Archive project and conform many current strategies, in the DARELUX project, the original bit streams are always saved with the converted ones. [7]

For our digital archive there is no fundamental difference between conversion and emulation. With emulation, the original bit stream is converted by one or more emulators. Conversion just means that the bit stream that results from the emulation is stored. Using emulation the conversion process is executed once the data is requested ('on the fly'), as with conversion the execution is done before and the result is stored. The three conversion strategies indicated above can be used with either conversion or emulation.

3 Data-sets and the library

Though the importance of data-sets is easily acknowledged by researchers, managing data-sets for a longer period and implementing quality assurance to them is not self evident. In the co-operation between the library and the researcher we found that:

- Data-sets are easily orphaned because research (projects) do not provide a solid base for an archive (for medium and long term storage)
- The quality of the data is dramatically improved by an structured ingest procedures
- The re-use of data-sets is largely depending on the ease of access

At the same time we observed that Libraries have been taking care of these issues for a long time be it based on traditional publications.

It is from this insight that we propose that:

- Every research project that deals with (measurement) data-sets should be made conscious about the short, medium and long term availability of their data
- If applicable, research project should allocate budget and time to facilitate short, medium and long term availability of (measurement) their data-sets
- The library should advertise it's ability and willingness to co-operate in such projects during the project (and so improve the quality and support the short term availability of the data-sets)
- The library should take care of medium and long term availability of the (measurement) data-sets

It might, on the other hand not be self evident for libraries to play a role as a data repository manager. Libraries tend to be traditional in their perception of intellectual output and consider the finished product (the publication) as the focus of their business. But since the scientific process is changing rapidly, adding more dynamics into it, libraries should, at least, consider whether they could play a role here.

On the other hand, already scientific data-centres have come into existence, and almost none in relation to the (traditional) library. So if the library foresees a role here for its future existence, it should work on advertising its willingness and ability to cooperate with researchers in this field while, at the same time, work on developing new knowledge and skills that are needed to become a professional partner for the researcher.

4 References

[1] DARELUX Data Archiving River Environment LUXemburg http://www.library.tudelft.nl/darelux/

[2] http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf

[3] Klump et al. : Data centres and their role in publication and access of data Ensuring long-term preservation and adding Value to Scietific and Technical data, *Royal Society of Edinburgh*, (November 2005)

[4] http://www.icdp-online.de/contenido/std-doi/front_content.php

[5] R. Dekker, K. van der Meer, E. Dürr. Digital preservation: The findings of the e-Archive project, N. Guimarães, P. Isaias (Ed): *Applied Computing* (2004). IADIS Press Lisbon, (2004) ISBN 972-98947-3-6

[6] R. Dekker, K. van der Meer and E. Dürr. Digital Preservation: The findings of the e-Archive project.. (2003). http://www.phys.uu.nl/~durr/EarchiveSite/publications/e-Archivefindingsfinal13.pdf

[7] R. Dekker et al.: An electronic archive for academic communities. In: P. Isaias (Ed.): *Proceedings of New Developments in Digital Libraries (NDDL)*, pp. 1-11,(2002). ECEIS Press, Setubal (Portugal), (2002).